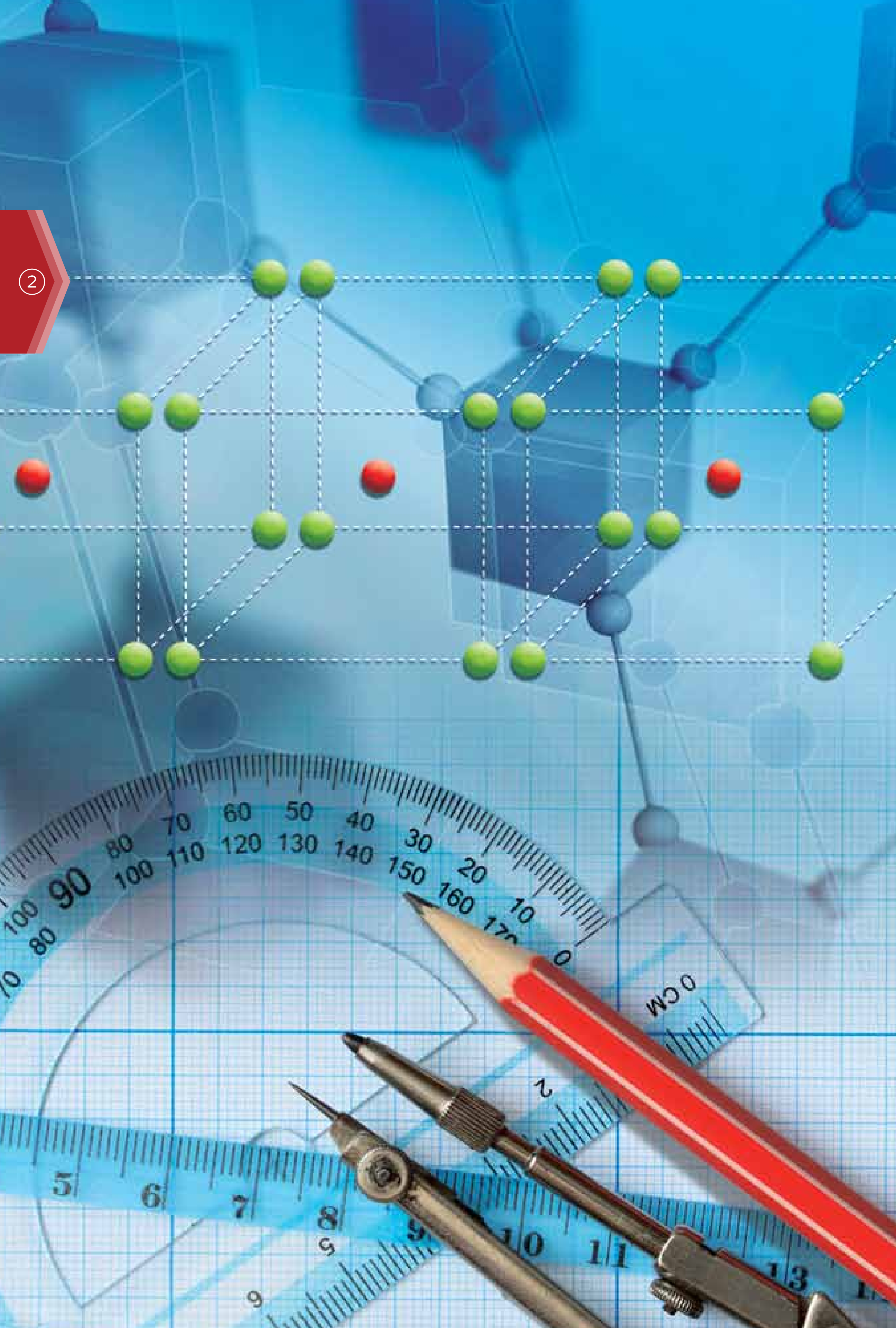
2


80 70 60 50 40 30 20 10 0
90 100 110 120 130 140 150 160 170
100 80 70
0 CM
2
5 6 7 8 9 10 11 13

# Report Documentation Page	*Form Approved* *OMB No. 0704-0188*

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **OCT 2014**	2. REPORT TYPE	3. DATES COVERED **00-00-2014 to 00-00-2014**

4. TITLE AND SUBTITLE **Improving Statistical Rigor in Defense Test and Evaluation: Use of Tolerance Intervals in Designed Experiments**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Defense Acquisition University,9820 Belvoir Road Ste 3,Fort Belvoir,VA,22060-9910**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S)
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**22**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

Keywords: *Test and Evaluation, Scientific Test and Analysis Techniques, Design of Experiments, Tolerance Intervals, Statistics*

Improving Statistical Rigor in Defense Test and Evaluation: *Use of Tolerance Intervals in Designed Experiments*

Alethea Rucker

Leveraging the use of statistical methods is critical in providing defensible test data to the Department of Defense Test and Evaluation (T&E) enterprise. This article investigates statistical tolerance intervals in designed experiments for the T&E technical community. Tolerance intervals are scarcely discussed in extant literature as compared to confidence/prediction intervals. The lesser known tolerance intervals can ensure a proportion of the population is captured in the design space, and have the ability to map the design space where factors can be reliably tested. Further, the article investigates several two-sided approximate tolerance factors estimated by Monte Carlo simulation and compares them to the exact method. Finally, the applicability of tolerance intervals to the defense T&E community is presented using a simple case study.

« *Image designed by Diane Fleischer*

In the FY 2012 Annual Report from the Director, Operational Test and Evaluation (DOT&E), the director identified two areas as requiring further improvement to move toward institutionalizing statistical rigor: (a) "execution of testing in accordance with the planned test design" and (b) "analysis of test data using advanced statistical methods commensurate with test designs developed using DOE [Design of Experiments]" (Gilmore, 2012a, p. v). The report further states that current data analysis is "limited to reporting a single average (mean) of the performance across all the test conditions" (p. v). In doing so, efficiencies achieved through meticulous test planning and design are discarded. Realizing the need for increased rigor, a Defense Science of Test Research Consortium was formed in 2011, partnered with Arizona State University, Virginia Polytechnic Institute and State University, Naval Postgraduate School, and the Air Force Institute of Technology (AFIT). The consortium's overall research goal is to support the incorporation of advanced statistical rigor and mathematical foundations into the test enterprise (AFIT, 2012). Research largely focuses on improved experimental design and statistical theory (Freeman, Ryan, Kensler, Dickinson & Vining, 2013; Haase, Hill, & Hodson, 2011; Hill, Gutman, Chambal, & Kitchen, 2013; Johnson, Hutto, Simpson, & Montgomery, 2012). The research of tolerance intervals in designed experiments has yet to be fully discussed. This article continues the research dialogue and adds to the body of knowledge of tolerance interval literature in defense testing, particularly the Scientific Test and Analysis Techniques (STAT) implementation effort. Further, this article aims to assist primarily test and evaluation (T&E) practitioners such as engineers, analysts, and test project/program managers in understanding how the use of statistics can greatly improve the quality of results in the decision-making process and improve credibility through objective data.

Purpose

The purpose of this article is twofold. First, tolerance intervals are rarely discussed in extant literature as having application to the defense T&E community. This research closes the gap by exploring the applicability of tolerance intervals in designed experiments. An attempt to use tolerance intervals in defense testing was investigated by the National Research Council (NRC) of the National Academies on testing body armor materials. Their recently published work (NRC, 2012) recommended use of statistical tolerance bounds, but their examples were confined to single, normally distributed samples, and did not take into account the design structure. Second, increased statistical rigor is needed in defense testing and analysis due to the complexity and challenges in testing a defense weapon system. Recently, the use of STAT has gained traction within the Department of Defense (DoD) T&E community (DoD, 2012; Gilmore, 2010; Operational Test Agencies, 2009). Albeit gradual, the defense community is leveraging the long-spanning, rich history of statistical methods in industry and replacing the budget-driven test events, combat scenarios, and one-factor-at-a-time approach with a statisticaly rigorous approach to test design using DOE (Johnson et al., 2012). Though current guidance and emphasis on the use of designed experiments in test plans is sufficient in explaining test planning and design, it falls short of providing specific guidance on test analysis and reporting to the decision makers, who ultimately decide on whether to field the weapon system to the warfighter. In these resource-constrained times, providing the T&E community defensible and objective test data to enable risk management for leadership during system development, procurement, and operation is imperative.

Increased statistical rigor is needed in defense testing and analysis due to the complexity and challenges in testing a defense weapon system.

What Are Tolerance Intervals?

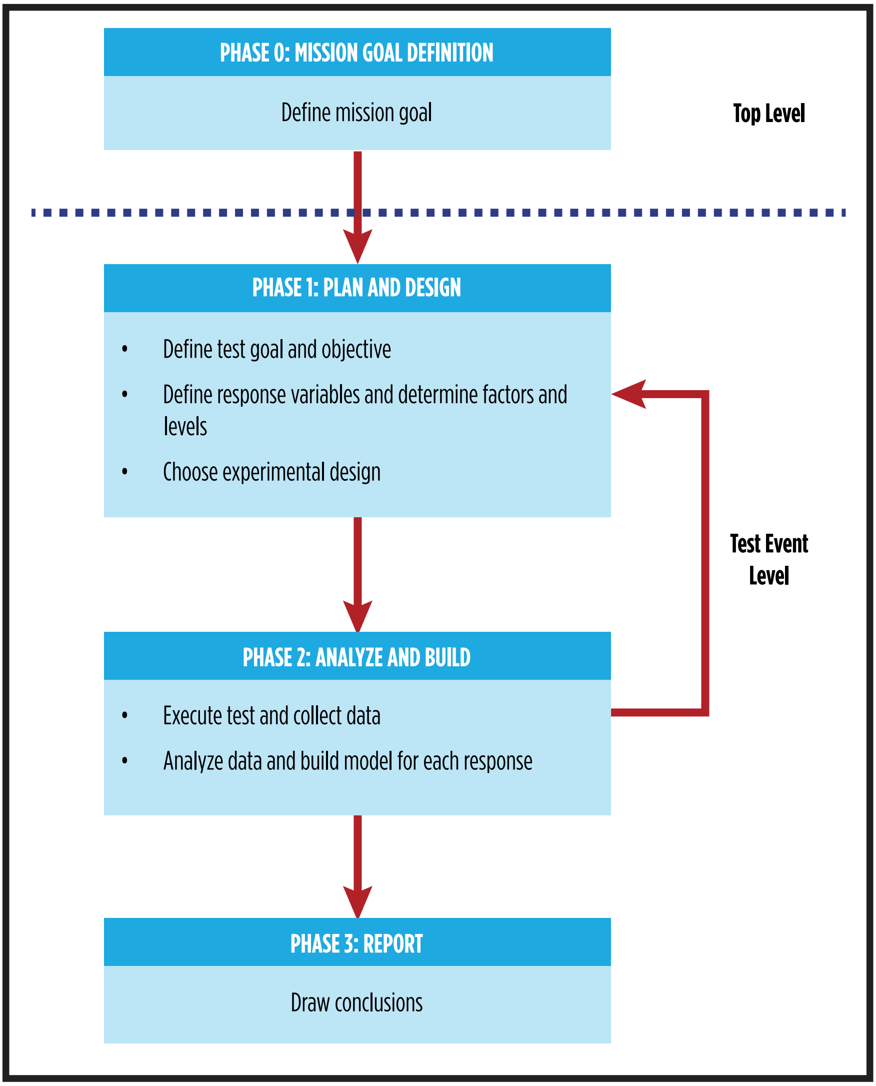
The importance of tolerance intervals has long been recognized (Wilks, 1941, 1942), with wide applicability to areas such as manufacturing, pharmaceutical, quality control, engineering, and material science commonly referred to as A- and B-basis allowables. In general, tolerance intervals capture a fixed proportion of population (p) with a given confidence level (1-α). Confidence intervals are the most commonly used statistical interval method focused on parameters such as mean and/or standard deviation, while prediction intervals consider the prediction of individual responses. Prediction intervals are useful only if the sample on which the interval is based represents the population, but if the population changes over time, then the prediction interval is useless (Vining, 1997). In other practical instances, the proportion of population, rather than mean is of interest, rendering tolerance intervals more appropriately applied in those situations. A tolerance interval allows us to make statements surrounding the distribution rather than the predicted individual responses, such as: "We are 95 percent confident that at least 90 percent of the population distribution will lie within the specified interval." Unfortunately, tolerance intervals are the least discussed interval in extant literature. Jensen (2009) attributes this to the difficulty of computation and lack of statistical software packages that readily offer tolerance intervals. De Gryze, Langhans, and Vandebroek (2007) indicated practical guidelines to calculate and use tolerance intervals in real-world applications are currently absent and that for even the simplest regression model, tolerance intervals are lacking.

The relevant interval in some situations in defense testing such as body armor testing should be one that states a specified proportion of population that falls above or below some threshold limit versus merely reporting the mean of the response. Many researchers have conducted studies in the construction of tolerance limits for normal distribution; early works by Wilks (1941, 1942), Wald (1943), and Wald and Wolfowitz (1946) are widely available in the literature. Exact methods for one-sided and two-sided tolerance regions have been researched for the normal distribution. Tolerance intervals for linear regression models were first introduced in the seminal paper by Wallis (1951). Wallis extended the previous work of Wald and Wolfowitz (1946) for a normally distributed sample to a linear regression model. Since then, researchers have extended Wallis's work to multiple and multivariate tolerance intervals (Krishnamoorthy & Mondal, 2008; Lee & Mathew, 2004). The continued research in this field has allowed T&E practitioners to expand their analysis and evaluation to include multiple responses such as time to acquire a target and miss distance. Not discussed in this article are Bayesian methods that incorporate a priori information and are useful in rolling up and including developmental test data and/or subject matter expert opinions. For a review of the statistical tolerance region, including Bayesian tolerance intervals, see Krishnamoorthy and Mathew (2009).

Design of Experiments Framework

A simplified DOE framework (Figure 1) is proposed for use in the DoD T&E community and later applied to the case study. Another suggested framework is the Plan, Design, Execute, Analyze model developed by Eglin Air Force Base, Florida.

FIGURE 1. GENERALIZED DOE FRAMEWORK


PHASE 0: MISSION GOAL DEFINITION
Define mission goal
Top Level
PHASE 1: PLAN AND DESIGN
Define test goal and objective
Define response variables and determine factors and levels
Choose experimental design
Test Event Level
PHASE 2: ANALYZE AND BUILD
Execute test and collect data
Analyze data and build model for each response
PHASE 3: REPORT
Draw conclusions

Phase 0: Mission Goal Definition

This step is accomplished at the top level and early in the acquisition cycle, where the mission statement and objectives (Critical Operational Issues, or COIs) are clearly defined for the program. COIs answer the question, "What capability will the system provide?" A hierarchy can serve as a catalyst for generating discussion about the identification of factors, levels, and responses for the proposed tests. This is accomplished by the test team—a T&E Working-level Integrated Program Team (WIPT). The T&E WIPT membership should include all stakeholder organizations from the developmental test and operational test communities. Test membership should include, but is not limited to, the program manager, operators, subject matter experts, program analysts, testers, and requirements representative. Generation of goals, objectives, factors, levels, and responses should be an exhaustive process so no input and output variables are left out. Therefore, continuously including these members upfront is critical to improving test outcomes.

Phase 1: Plan and Design

Test goal and objective. Every good experimental design begins with a clear, concise goal and objective that is well understood by all parties before test planning. The right kind of questions leading to development of quantifiable terms (responses) need to be articulated for effective test execution and data collection. The right quantitative metrics are essential for developing a good test design; poorly chosen or ill-defined measures can lead to unnecessary costs or ambiguous test results (Gilmore, 2012b). Continuous metrics, such as detection range, enable the most efficient use of resources and provide the most information. On the other hand, binary metrics, such as pass or fail, hit or miss, offer less information to testers and can increase test resource requirements.

Response variables, factors, and levels. Selection of a response variable, continuous or discrete, should be carefully considered to minimize risk in running into a Type I (α) or Type II (β) error. Responses are Key Performance Parameters, Measures of Effectiveness, Measures of Suitability, Critical Technical Parameters, Key System Attributes, and/or Measures of Performance that are documented and traced to the requirements document (Gilmore, 2012b). In current DoD test planning, the statistical measures of merit—power ($1-\beta$) and confidence level ($1-\alpha$)—must be documented in the Test and Evaluation Master Plan (TEMP; see Gilmore, 2012b). The Type I error (α) is the probability of declaring a factor is affecting the response when in reality it does not. This percent value is typically agreed upon by the decision maker based on inputs from the T&E WIPT. The quantity expresses the decision makers' risk tolerance for making a wrong decision based on limited test data (Freeman, Glaeser, & Rucker, 2011). A Type II error (β) is the probability of declaring a factor does not affect the response, when in reality it does. Power ($1-\beta$) is the likelihood of not making the β error and the ability to detect differences. This is set by the test team during test planning. In general, the confidence levels are set between 80% and 95% (α = 0.20 to 0.05), and the power for a signal-to-noise S:N = 1.0 should be above 80% (Department of the Army, 2012). Both types of errors must be well understood and explained to the decision maker due to the unintended programmatic consequences that might result from a lack of understanding. Defining factors (independent variables) is no trivial matter and must be determined by the entire test team. The factors define the operational environment of the system. Some proven effective brainstorming methods to aid the process are fishbone diagrams (also known as Ishikawa Diagrams) and process flow diagrams. Note that it is better to include more factors than preclude factors that might be significant (Telford, 2007). Levels are the specified values of the factors, and the general recommendation is to consider two to three levels for each factor (Freeman et al., 2011).

Test design. The test design is constructed after the factor, levels, and responses are identified by the test team. Decision trade-offs between risk and costs are made at this stage with assistance from the test team, especially when test resources are limited. The choice of design involves consideration of sample size, selection of a suitable run order for the trials, and whether blocking or other randomization restrictions exist (Montgomery, 2001). Depending on the test conditions, copious sources of noise might be present and must be considered in the test design. Also, test resources and programmatic constraints may prohibit common designs. Consider the basic principles of DOE when designing a test: randomization, replication, and blocking. Randomization is the underlying foundation of the use of statistical methods. It reduces the likelihood of introducing bias to the experiment by randomizing the effect of uncontrolled variables, such as unplanned weather effects. Replication of test points allows for estimation of system variability and test procedure error. Blocking provides another way to address variability and improves the power to detect a factor effect (Freeman et al., 2011). Coleman and Montgomery (1993) provide guidelines in the preexperimental planning phases to assist with designing and conducting an experiment. Some other papers useful in explaining experimental planning and design include Hunter (1977) and Montgomery (2005).

Phase 2: Analyze and Build

If all steps leading to this phase are properly and thoroughly planned, then the test is well defined. However, no test execution goes as planned due to nuisance factors and noise that exist such as weather and/or data processing. Data are collected at this phase and analyzed by the test team. A mathematical model is created for each response variable by mapping a response surface over the region of interest (operational range) so that the effect of factors on that response can be studied (Johnson et al., 2012). The analysis will result in generating statistically defensible models that inform the decision maker.

Phase 3: Report

The test team should draw conclusions based on information extracted from test data. Appropriate scientific test and analysis techniques should be employed so that senior leaders can make an informed decision backed by defensible data.

Brief Review of Two-Sided Tolerance Intervals

In this section, two-sided approximation tolerance intervals are considered. To describe a general two-sided tolerance interval form, let x_1, x_2, .., x_n be values of a random sample X_1, X_2, .., X_n of size *n* from a normal distribution with mean μ and variance σ^2 where:

$$X \sim N(\mu, \sigma^2)$$

The 100(P)% two-sided tolerance interval with confidence 100(1-α)% is of the form $\bar{x} \pm k_2 s$ for which the following applies:

$$Pr[P(\bar{x} - k_2 s < X < \bar{x} + k_2 s) \geq P] = 1 - \alpha$$

where $\bar{x}$ is the sample mean, k_2 is a constant multiplier, *s* is the sample standard deviation, 1-α is the confidence level associated with the interval, and *P* is the proportion of distribution covered by the interval, referred to as coverage. To describe the two-sided *regression* tolerance interval, let's consider the general structure for a regression model:

$$y_i = \beta_0 + \beta x_i + \varepsilon_i, i = 1,\ldots,n$$

where y_i is the p *x* 1 response vector, x_i is the known m *x* 1 factor variable vector, β_0 is the p *x* 1 intercept vector, β is unknown p *x* m regression parameter vector, and ε_i is assumed to be a vector of independent, normally distributed error terms, each with mean zero and variance σ^2. To estimate β, least squares regression is applied based on a set of n observations. The predicted mean response would then be of the form:

$$\hat{y} = x\hat{\beta}$$

Suppose for any known value of factors $x = x_i$ with a corresponding fitted value $\hat{y}_i$, the 100(P)% two-sided tolerance interval with confidence 100(1-α)% is of the form:

$$\hat{y}_i \pm k_{2,i} s$$

where k is the tolerance factor and s^2 is the residual mean square error based on degrees of freedom.

Monte Carlo Evaluation of Tolerance Intervals

The first approximation considered was proposed by Howe (1969). Howe introduced an approximate factor for a two-sided tolerance interval for a normally distributed population given as:

$$\mathrm{k}_2 = \sqrt{\frac{df(1+\frac{1}{n})\mathrm{z}^2_{(1-P)/2}}{\mathrm{x}^2_{1-\alpha,\, df}}}$$

where is $x^2_{1\text{-}\alpha,\, df}$ the αth percentile of the chi-square distribution with df, degrees of freedom, n is the sample size, $df = n - m$ (number of independent random samples) is degrees of freedom defined as the number of values that are free to vary, and $\mathrm{z}_{(1\text{-}P)/2}$ is the pth percentile of the standard normal distribution.

The second approximation was proposed by Zorn, Gibbons, and Sonzogni (1997). They introduced a weighted tolerance interval for estimating detection and quantification limits in the chemical field. Leveraging the earlier work of Lieberman and Miller (1963) in developing simultaneous tolerance intervals for linear regression, they translated it to a nonsimultaneous case. The two-sided approximation (TI_2) would result in:

$$TI_2 \approx \widehat{y_0} \pm s\left[\mathrm{t}_{\frac{\alpha}{2},df}\sqrt{x_0^T(X^TX)^{-1}} + \Phi^{-1}(P)\sqrt{\frac{df}{\chi^2(1-\frac{\alpha}{2},df)}}\right]$$

where x_0 is a point in the design space, X is the design matrix of the regression model, $\Phi^{-1}(P)$ is the inverse cumulative normal distribution, and $t_{\frac{a}{2}}$,*df* is the Student's *t*-inverse cumulative distribution function using degrees of freedom for the corresponding confidence.

The third approximation is credited to De Gryze et al. (2007) when they proposed taking α in both χ2 (*df*) and *t*(*df*) quantiles, thus resulting in the approximation below:

$$TI_2 \approx \widehat{y_0} \pm s\left[\mathrm{t}_{\alpha,\,df}\sqrt{x_0^T\,(X^TX)^{-1}_{x_0}} + \Phi^{-1}(P)\sqrt{\frac{df}{\chi^2(1-\alpha,\,df)}}\right]$$

where x_0 is a point in the design space, X is the design matrix of the regression model, $\Phi^{-1}(P)$ is the inverse cumulative normal distribution, and $t_{\frac{\alpha}{2}}$,*df*) is the Student's *t*-inverse cumulative distribution function using degrees of freedom for the corresponding confidence.

The final method introduced is the exact two-sided tolerance interval due to Krishnamoorthy and Mathew (2009). The k is the solution of the integral equation:

$$2m\int_0^\infty P\left(1 - F\chi^2_{df}\left(\frac{df}{k^2}\chi^2_{1;P}\,(d^2z^2)\right)\right)(2\Phi(z) - 1)^{\mathrm{m}-1}\phi(z) = 1 - \alpha$$

where *df* is the degrees of freedom, $d^2 = x'(X'X)^{-1}\,x$ where X is the design matrix of the linear regression model, m is the number of independent random samples (factors), $\Phi(z)$ is the cumulative distribution function, $\phi(\mathrm{z})$ is the probability density function, $F\chi^2_{\mathrm{df}}$ is the cumulative distribution function of a chi-square distribution with *df* degrees of freedom. Detailed derivation of the exact equation can be found in Howe (1969), equations 1.2.3, 1.2.4, 2.5.7, and 2.5.8) and Witkovsky (2013). This article employs the MATLAB tolerance package developed by Witkovsky (2009) using the Gauss-Kronrod quadratic formulae for integration.

The following Monte-Carlo simulation algorithm is applied to approximations by De Gryze et al. (2007), Howe (1969), and Zorn et al. (1997).

1. Simulate 500 design points (x_0) within the test space uniformly distributed throughout the design space.

2. For a given 1-α, p, compute tolerance interval multiplier at x_0 (design point in the test space).

3. Since the tolerance interval multiplier is a function of the position in the design space, take the average tolerance interval multiplier value (based on the Monte Carlo simulation sample of 500 points previously mentioned) and multiply by $t_{\alpha,df}$ to determine the tolerance factor.

Next, the following are computed and compared: (a) the approximate factor by De Gryze et al. (2007), (b) the approximate factor by Zorn et al. (1997), (c) the approximate factor by Howe (1969), and (d) the exact tolerance factor. Figure 2 depicts the comparison case of p = 0.99 and 1-a = {0.90. 0.95, 0.99}.

FIGURE 2. P = 0.99, 1-α = {0.90, 0.95, 0.99}

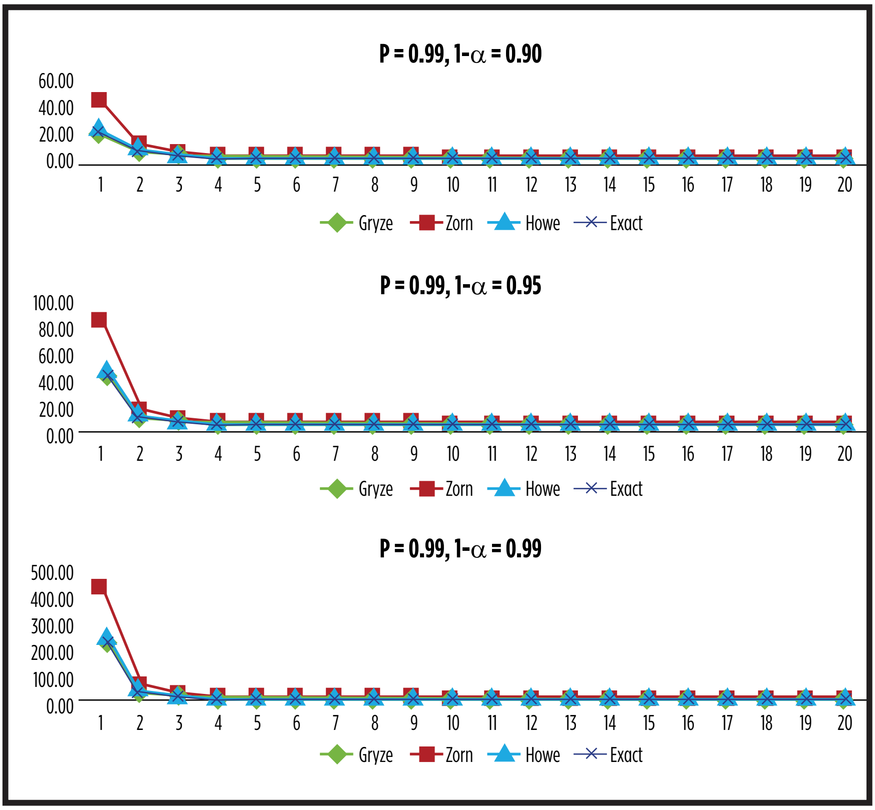


Figure 2 shows that all tolerance factors decrease as degrees of freedom increase. Immediately apparent is that the Zorn et al. (1997) approximation is the most conservative, especially for smaller degrees of freedom. All numerical results are slightly above the exact method, but the Howe (1969) approximation is slightly lower with larger degrees of freedom. The two-sided tolerance intervals performed well for degrees of freedom 4 and above. The approximated values performed well against the exact method, but it can be noted that Zorn's method would require a larger sample size. In general, to cover a multifactor region requires a wider tolerance region compared to the normal sample. So why investigate approximation methods when an exact method is available? An exact method calculation can be extremely complex and is rarely if ever available on statistical software packages. It can also be thought of as costly, given the difficulty and time involved in obtaining an exact solution. Therefore, approximation methods are generally preferred, but their accuracy is seldom confirmed. If an approximation needs to be used, the author recommends the De Gryze et al. (2007) proposed approximate method as a statistical test analysis method commensurate with designed experiments. The appeal of the De Gryze et al. (2007) method stems from the fact that this method takes into account the design structure and variance, is easier to compute, and is comparable to the exact method.

Use of Tolerance Intervals in Designed Experiment Case Study

This section will apply the two-sided tolerance interval to a designed experiment using a notional case study. The case study used throughout the article is for academic purposes and is by no means representative of any existing weapon employed by the DoD. Some aspects of the case study have been simplified for educational purposes.

Phase 1: Plan and Design

Objective. The objective of the experiment is to characterize the performance of a new and old air-to-ground missile.

Response variables, factors, and levels. Figure 3 and Table 1 show the factors and levels generated by the test team during the test design planning phase. The response variables are miss distance and impact velocity error.

FIGURE 3. FACTORS AND LEVELS GENERATED BY TEST TEAM DURING TEST DESIGN PLANNING PHASE

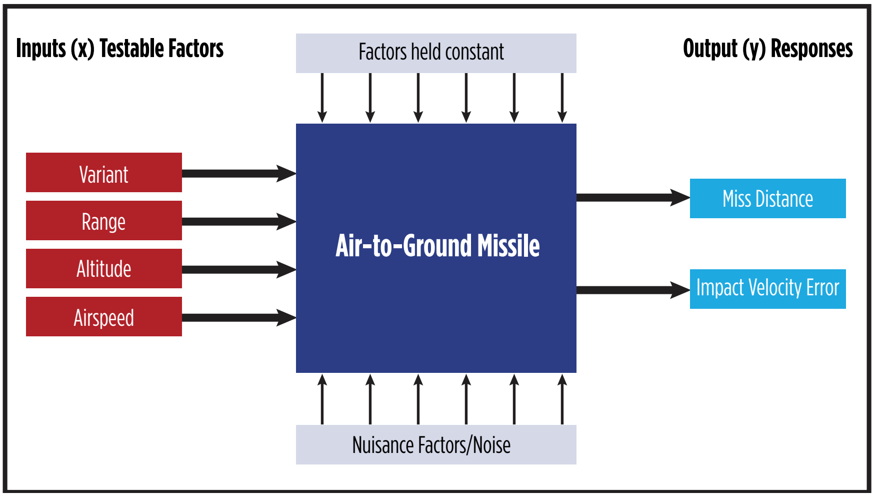


TABLE 1. FACTORS AND LEVELS

	Factor	Levels
A	Variant	0 (Legacy), 1 (New)
B	Range	-1, 1
C	Altitude	25, 35
D	Airspeed	0.85, 0.95

Test design. A 2^4 factorial test design, 16 runs were selected for this case study.

Phase 2: Analyze and Build

Test execution. Suppose the test team executed the test, and Table 2 reflects the results collected.

Model building. A regression model was built and analyzed. The overall response models were significant; range and airspeed were the two most important factors in characterizing the air-to-ground missile performance, and there was no statistical difference between the legacy and new variant across the operational envelope. However, for this research, analysis will be limited to the tolerance interval computation under high-airspeed and high-range conditions.

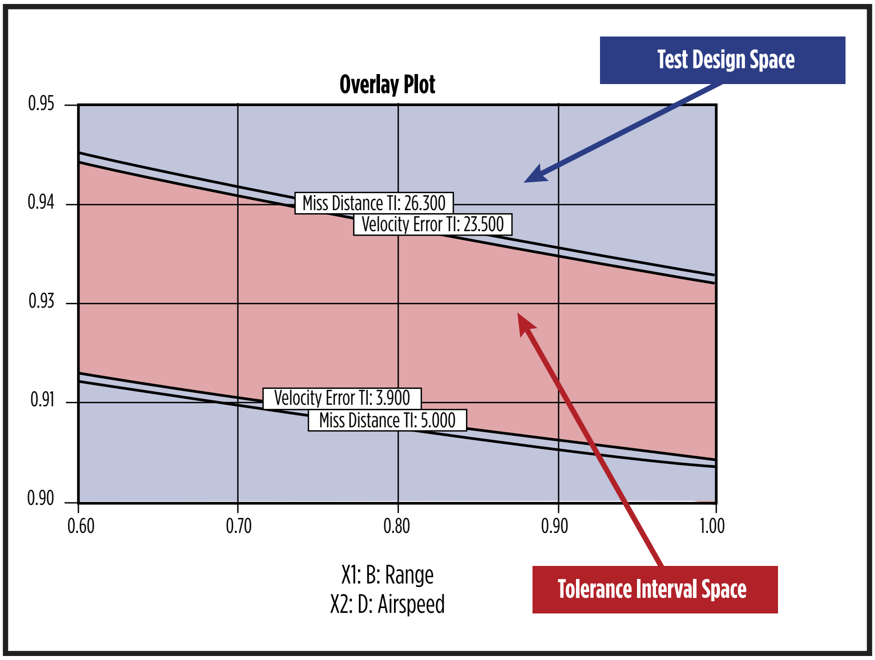
TABLE 2. AIR-TO-GROUND CASE STUDY TEST DESIGN AND RESULTS

Run	A: Variant	B: Range	C: Altitude	D: Airspeed	Miss Distance	Impact Velocity Error
1	1	-1	25	0.95	3.44	1.76
2	0	1	25	0.95	20.09	18.96
3	0	1	35	0.85	5.63	3.4
4	1	-1	35	0.95	8.58	6.71
5	1	-1	35	0.85	1.14	0.76
6	1	1	35	0.95	20.81	18.46
7	0	-1	25	0.85	4.65	2.83
8	1	1	25	0.85	4.45	2.49
9	1	1	25	0.95	19.9	17.51
10	1	1	35	0.85	5.44	3.86
11	0	1	35	0.95	22.47	20.35
12	0	-1	35	0.85	3.55	1.61
13	1	-1	25	0.85	3.04	1.38
14	0	-1	35	0.95	13.76	11.45
15	0	-1	25	0.95	7.58	5.39
16	0	1	25	0.85	5.23	3.7

Phase 3: Report

Based on the models generated, the following values were obtained: mean miss distance response value = 9.36 feet, miss distance standard deviation = 2.33 feet, mean impact velocity error response value = 7.54 feet/s, impact velocity error standard deviation = 2.15 feet/s, and degrees of freedom = 15. Suppose the test team has selected a, confidence = 0.05 (95%) and p, proportion of population = 99%. Referring to the De Gryze et al. (2007) approximation, the test team was able to report "with 95% confidence, at least 99% of the miss distance population will be between 5.0 to 26.3 ft., and at least 99% of the velocity error population will be between 3.9 ft./s to 23.5 ft./s under the specified condition." Now the test team overlaid the two most important factors—range and airspeed—and bounded the "test design space" with tolerance intervals obtained previously for condition 1 (Figure 4). From this plot, the team can easily extract the "sweet spot," "operating window," or in this case the "tolerance interval space" where they can ascertain with a specified confidence that at least 99 percent of both responses would be found, under the specified conditions.

FIGURE 4. RANGE AND AIRSPEED OVERLAY PLOT



Further, the test team investigated how confidence and tolerance intervals compared. The 95 percent confidence interval for miss distance mean and impact velocity error mean were found to be within 13.8 to 17.5 feet, and 12 to 15.4 feet/s, respectively. This means the "true" mean of the miss distance and impact velocity error measurements lies within these bounds. Oftentimes, we might not need to place bounds on the distribution parameters, but on the specified proportion of population instead, hence the appeal of tolerance intervals. The confidence interval may win the interval popular vote; however, the beauty of the tolerance interval lies in the fact it takes into account not only the sample size, but also the estimates of mean and standard deviation noise. Given the test data generated, the test team was able to narrow down and recommend a specific response interval where 99 percent of the population would lie at the factors identified under the specified conditions. This enabled the program manager to set an operational window where the air-to-ground missile would perform at its optimum for high airspeeds and high range. (Recall the case study is notional, but this is an illustration of the type of information that can be drawn.)

In this case study, a statistical tolerance interval ensured a defensible conclusion with a sound analytical basis, rather than simply stating the mean as criticized in the DOT&E *FY 2012 Annual Report*. Through the combined use of DOE, regression analysis, and tolerance intervals, T&E practitioners are able to frame the operating window with some confidence and have the ability to map out the test space where factors can be reliably tested. This is a significant improvement over simply stating a single average across all test conditions, and it allows us to extract more information from limited resources and test events. The efficiencies obtained through the meticulous planning using DOE principles were retained. An advanced statistical analysis that complements DOE proved capable of defining an operating window with some certainty and well-understood risks where the air-to-ground missile can be adequately operated. Understanding the appropriate use of statistical analysis technique is imperative and does matter; for example, interaction effects need to be considered and a simple one-way analysis of variance or use of average value might ignore or hide the interaction between main effects. Therefore, the research into suitable advanced statistical analysis methods commensurate with DOE needs to continue.

Through the combined use of DOE, regression analysis, and tolerance intervals, T&E practitioners are able to frame the operating window with some confidence, and have the ability to map out the test space where factors can be reliably tested.

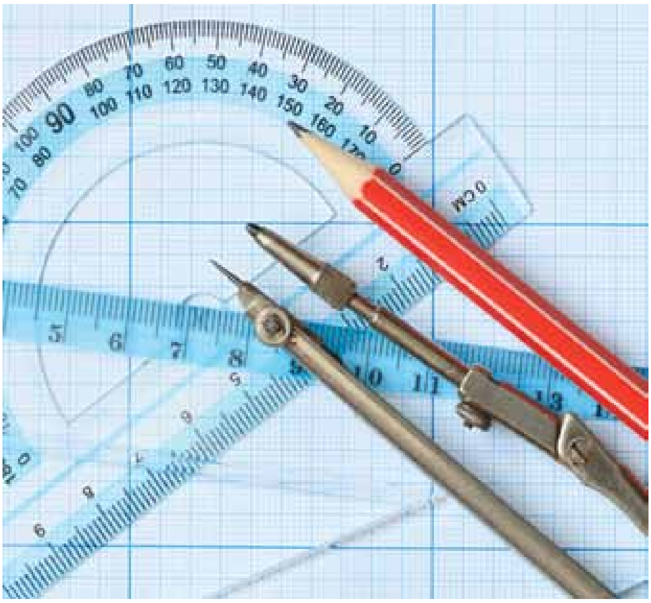
Limitations and Future Research

In general, tolerance intervals offer a more useful means to assure, with some confidence, that a fixed proportion of the systems' performance over the design space falls within a specified interval. The analysis method reaps the benefits of a designed experiment and employs statistical techniques that are commensurate with DOE. This research, however, does have limitations, indicating a need for further discussion and research. Future research should include qualitative metrics, such as categorical factors. In addition, other tolerance intervals such as nonparametric regression tolerance intervals should be investigated for future use in the defense test community (see Young, 2010, for other intervals). One-sided regression tolerance intervals for defense testing should be presented and compared using the proposed Monte Carlo simulation algorithm; the calculation is generally simpler than the two-sided case. When exact methods are not available, the author recommends using the approximate methods mentioned in this article that are best suited for multiple regression models. Be forewarned that the use of tolerance intervals may require a larger sample size; for this reason and to properly size your experiment, the author also recommends investigating the use of tolerance intervals in test planning (see Whitcomb & Anderson, 2011, for examples).

Recommendations and Conclusions

The defense T&E community has progressed in its efforts to advance statistical rigor within the community over the past 3 years; however, some areas still need improvement. One area is to improve interaction between necessary stakeholder organizations and the T&E community. All organizations that have an impact and/or influence on the program's T&E planning, execution, and assessment need to be engaged in the T&E WIPT as early as possible. Another area would be to increase education and training on the use of STAT for all stakeholders, and this means going above and beyond what confidence intervals provide. Finally, best practices, lessons learned, and research need to be continuously published and readily available to the T&E community.

In these resource-constrained times, every dollar spent on defense must count. As the DoD moves toward generating defensible data through the use of DOE for test designs and institutionalizing statistical rigor within the T&E community, it seems logical to employ advanced statistical analysis methods that reap the benefits afforded by DOE to generate efficiencies. Rigor should not end with the test design, and solid analytical evidence needs to be presented all the way through test reporting. The literature to date does not adequately address the appropriate use of defensible data developed through improved design methods, nor does it propose a statistical analysis, such as tolerance intervals, commensurate with test designs developed using DOE for the defense community. This article fills that gap by introducing the applicability of tolerance intervals as an analysis technique in a designed experiment and by comparing several two-sided approximate tolerance factors estimated by Monte Carlo simulation to the exact method. Further, this article provides a recommendation of the most appropriate tolerance interval and its applicability to the defense T&E community using a simple case study. This analytical method provides a meaningful objective way to add rigor to an otherwise subjective assessment, extracts more information to state how the system will perform in the operational conditions, and serves as a quantitative decision aid to our senior leaders.



Author Biography

Dr. Alethea Rucker is a senior aerospace engineer at the Defense Technology Security Administration, Technology Directorate, Pentagon, Washington, DC. Prior to this, she was the deputy division chief, Assistant Secretary of the Air Force (Acquisition), Space Programs. She has held flight test and dynamics positions at Air Force Test Center, Northrop Grumman Corporation, and Parker Hannifin (Aerospace). She holds a BS and MS in Aeronautical and Astronautical Engineering from Purdue University, and a PhD in Systems Engineering from The George Washington University.

(E-mail address: alethea.rucker.civ@mail.mil)

References

Air Force Institute of Technology. (2012). *Science of test research consortium: Year two final report* (Report No. AFIT/EN/TR-13-01). Wright-Patterson AFB, OH: Author.

Coleman, D. E., & Montgomery, D. C. (1993). A systematic approach to planning for a designed industrial experiment. *Technometrics, 35*(1), 1–12.

De Gryze, S., Langhans, I., & Vandebroek, M. (2007). Using the correct intervals for prediction: A tutorial on tolerance intervals for ordinary least-squares regression. *Chemometrics and Intelligent Laboratory Systems, 87*(2), 147–154.

Department of the Army. (2012). *Implementation of Design of Experiments (DOE) techniques for test and evaluation (T&E)* (ATEC Policy Bulletin 3-12). Aberdeen Proving Ground, MD: Author.

Department of Defense. (2012). *Scientific test and analysis techniques in test and evaluation implementation plan.* Washington, DC: Author.

Freeman, L., Glaeser, K., & Rucker, A. (2011). Use of statistically designed experiments to inform decisions in a resource constrained environment. *International Test and Evaluation Journal, 32*(3), 267–276.

Freeman, L. J., Ryan, A. G., Kensler, J. L., Dickinson, R. M., & Vining, G. G. (2013). A tutorial on the planning of experiments. *Quality Engineering, 25*(4), 315–332.

Gilmore, J. M. (2010). *Guidance on the use of Design of Experiments (DOE) in operational test and evaluation.* Washington, DC: Office of the Director, Operational Test and Evaluation.

Gilmore, J. M. (2012a). *FY 2012 annual report.* Washington, DC: Office of the Director, Operational Test and Evaluation.

Gilmore, J. M. (2012b). *TEMP guidebook.* Washington, DC: Office of the Director, Operational Test and Evaluation.

Haase, C. L., Hill, R. R., & Hodson, D. (2011). Using statistical experimental design to realize LVC potential in T&E. *International Test and Evaluation Journal, 32*(3), 288–297.

Hill, R. R., Gutman, A. J., Chambal, S. P., & Kitchen, J. W. (2013). Acquisition and testing, DT/OT testing: The need for two-parameter requirements. *Quality and Reliability Engineering International, 29*(5), 691–697.

Howe, W. G. (1969). Two-sided tolerance limits for normal populations—Some improvements. *Journal of the American Statistical Association, 64*(326), 610–620.

Hunter, W. G. (1977). Some ideas about teaching Design of Experiments with 25 examples of experiments conducted by students. *The American Statistician, 31*(1), 12–17.

Jensen, W. A. (2009). Approximations of tolerance intervals for normally distributed data. *Quality and Reliability Engineering International, 25*(5), 571–580.

Johnson, R. T., Hutto, G. T., Simpson, J. R., & Montgomery, D. C. (2012). Designed experiments for the defense community. *Quality Engineering, 24*(1), 60–79.

Krishnamoorthy, K., & Mathew, T. (2009). *Statistical tolerance regions: Theory, applications, and computation* (Vol. 744). Hoboken, NJ: John Wiley & Sons.

Krishnamoorthy, K., & Mondal, S. (2008). Tolerance factors in multiple and multivariate linear regressions. *Communications in Statistics—Simulation and Computation, 37*(3), 546–559.

Lee, Y. T., & Mathew, T. (2004). Tolerance regions in multivariate linear regression. *Journal of Statistical Planning and Inference, 126*(1), 253–271.

Lieberman, G. J., & Miller, R. G. (1963). Simultaneous tolerance intervals in regression. *Biometrika, 50*(1–2), 155–168.

Montgomery, D. C. (2001). *Design and analysis of experiments* (5th ed.). Hoboken, NJ: John Wiley & Sons.

Montgomery, D. C. (2005). *Design and analysis of experiments* (6th ed.). Hoboken, NJ: John Wiley & Sons.

National Research Council. (2012). *Testing of body armor materials: Phase III.* Washington, DC: The National Academies Press.

Operational Test Agencies. (2009). *Using design of experiments for operational test and evaluation.* Washington DC: Office of the Director, Operational Test and Evaluation/Air Force Operational Test and Evaluation Center.

Telford, J. K. (2007). A brief introduction to design of experiments. *Johns Hopkins APL Technical Digest, 27*(3), 224–232.

Vining, G. G. (1997). *Statistical methods for engineers*. Pacific Grove, CA: Duxbury Press.

Wald, A. (1943). An extension of Wilks' method for setting tolerance limits. *The Annals of Mathematical Statistics, 14*(1), 45–55.

Wald, A., & Wolfowitz, J. (1946). Tolerance limits for a normal distribution. *The Annals of Mathematical Statistics, 17*(2), 208–215.

Wallis, W. A. (1951). Tolerance intervals for linear regression. In J. Neyman (Ed.), *Second Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 43–51). Los Angeles and Berkeley: University of California Press.

Whitcomb, P. J., & Anderson, M. J. (2011, September). Using DOE with tolerance intervals to verify specifications. In *Proceedings of the 11th Annual Meeting of the European Network for Business and Industrial Statistics (ENBIS),* (pp. 4–8). Coimbra, Portugal: University of Coimbra.

Wilks, S. S. (1941). Determination of sample sizes for setting tolerance limits. *The Annals of Mathematical Statistics, 12*(1), 91–96.

Wilks, S. S. (1942). Statistical prediction with special reference to the problem of tolerance limits. *The Annals of Mathematical Statistics, 13*(4), 400–409.

Witkovsky, V. (2009). *Tolerance Factor.* MATLAB Central File Exchange [Online file exchange/library]. Retrieved from http://www.mathworks.com/matlabcentral/fileexchange/24135-tolerancefactor

Witkovsky, V. (2013). On the exact tolerance intervals for univariate normal distribution. In Proceedings of *Computer Data Analysis & Modeling*, Minsk, Belarus, September 10–14.

Young, D. S. (2010). Tolerance: An R package for estimating tolerance intervals. *Journal of Statistical Software, 36*(5), 1–39.

Zorn, M. E., Gibbons, R. D., & Sonzogni, W. C. (1997). Weighted least-squares approach to calculating limits of detection and quantification by modeling variability as a function of concentration. *Analytical Chemistry, 69*(15), 3069–3075.